Accepted for presentation at the AMIA Annual Symposium 2006
draft material -- please do not cite

Using WordNet to Improve the Mapping of Data Elements to UMLS for Data Sources Integration

Fleur Mougin[a], Anita Burgun[a], M.D.,Ph.D., Olivier Bodenreider[b], M.D.,Ph.D.
[a]EA 3888, IFR 140, Faculté de Médecine, Université de Rennes I, France
[b]National Library of Medicine, Bethesda, Maryland
{fleur.mougin,anita.burgun}@univ-rennes1.fr, olivier@nlm.nih.gov

Each biomedical system has its own way of naming the pieces of information it contains, i.e., of defining its data elements (DEs). Integrating DEs facilitates the integration of biomedical resources. However, the mapping of DEs to the UMLS is ambiguous in many cases, when any correspondence is found at all. We propose to evaluate the potential contribution of a more general terminology: WordNet. Our method is based on synonyms, definitions, and structural properties of the terminologies. We applied it to a set of 474 DEs extracted from eleven biomedical sources. We show that WordNet can improve the direct mapping of DEs to UMLS when used to validate and disambiguate UMLS direct mappings. WordNet can also help identify indirect mappings of DEs to the UMLS.

INTRODUCTION

Because most biomedical systems have been developed independently of each other, they do not have a common structure, nor do they share a common data dictionary or data elements (DEs). DEs can be defined as follows[1]:

- A named identifier of each of the entities and their attributes that are represented in a database.
- A basic unit of information built on standard structures having a unique meaning and distinct units or values.

Examples of DEs in the biomedical domain include `Gene Symbol` and `Pathology Name`. The corresponding value sets would be the set of gene symbols (e.g., in a given model organism) and a list of diseases, respectively.

In practice, the major barriers to data integration are the heterogeneity of database schemas and the disparity of DEs across systems. The general framework of this paper is the integration of DEs in order to enable the integration of biomedical resources.

In a previous study [1], we used the Unified Medical Language System® (UMLS®) [2] for mapping DEs coming from separate Web resources to a biomedical terminology in order to integrate them. Toward this end, we attempted to find an exact match and a normalized match, by using existing lexical tools [3]. Finally, when no match was found, an approximate match was attempted using MetaMap, a program allowing to map text to concepts in the Metathesaurus [4]. The output of this mapping consists of the list of Metathesaurus concepts for each DE, along with their semantic types, textual definition (when provided), synonymous terms, and ancestors.

The outcome of the mapping of DEs to the UMLS can be summarized as follows:

- *Unique match*. For example, the DE `Additional cdna sequence` is mapped to the concept “DNA, Complementary” by approximate match.
- *Multiple matches*. For instance, the DE `Protein` results in an exact match to three UMLS concepts: “Protein”, “Protein measurement”, and “Protein location”.
- *No match*. Some DEs are simply not mapped to any UMLS concepts, because they are not specific to the biomedical domain and need to be represented at a higher (more general) level. Examples of such DEs include `features`, `keywords`, and `domains`.

We propose to improve the mapping to the UMLS by using supplementary information. Our hypothesis is that general resources such as WordNet [5], an online lexical database of general English, could give a complementary coverage of the domain described by the studied DEs. Previous studies have underlined common characteristics existing between these two terminological resources [6], making it possible to align them automatically and accurately. More specifically, by exploiting the properties of WordNet (WN), we expect to improve the mapping of DEs to the UMLS in the following ways. In case of unique matches, WN would help validate the UMLS mappings. This can be especially useful when MetaMap resolves acronyms (e.g. cDNA, as illustrated above), which is often error-prone. For multiple matches, WN would contribute external information useful for disambiguating UMLS mappings. Finally, WN would help identify indirect mappings to the UMLS when no direct UMLS mapping was found.

The objectives of this study are to validate and disambiguate the direct mappings of DEs to the UMLS using information from WN. Additionally, we propose to identify indirect mappings to the UMLS (through WN) for those DEs for which no direct match was found.

[1] http://www.atis.org/tg2k/_data_element.html

MATERIALS

DATA ELEMENTS

Origin. Our test set consists of data elements extracted from eleven Web-accessible biomedical sources, selected to be representative of the different kinds of resources found in the biomedical domain. Some of them contain information about genes: GeneCards[2], Entrez Gene[3], Geneloc[4], Genew (the HGNC[5] database), and HGMD[6], others about proteins: Swiss-Prot[7], PDB[8], HPRD[9], Interpro[10] or diseases: OMIM[11]. Our application is not targeted to a particular model organism so we also included the resource MGI[12], which provides various kinds of information about mice.

Extracting data elements

Creating a set of query terms. We first assembled a set of biomedical terms to be used as query terms in the data sources under investigation. These terms were extracted manually from a reference resource in the domain of medical genetics: the Genetics Home Reference[13]. Our data set includes 100 terms such as gene symbols (e.g. HFE, BRCA1) and pathologies (e.g. hemochromatosis, breast cancer).

Querying data sources. Each of the eleven sources is queried automatically for each term. In practice, the procedure used to query the sources can be described as follows.

- Identifying the URL allowing to query it dynamically.
- Creating a set of 100 HTML pages corresponding to entries of the set of biomedical terms.
- Pre-processing each page by first eliminating the header and footer, which are common to HTML pages. In fact, many of the resources used in this study are Web interfaces to biological databases, automatically generated by program. Therefore, it is expected that most pages of a given resource share a common organization and presentation. We take advantage of this feature for identifying recurring terms throughout pages, which, we hypothesize, correspond to data elements.
- Selecting the terms (extracted from the different HTML pages) that are common to at least 75% of the HTML pages. This selection results in eliminating specific information (e.g., a given gene name) while keeping general information (e.g., the term "Gene Name").

Examples of data elements extracted from the source Genew are `Approved Symbol` and `Previous Names`.

Integrating data elements. The data elements (DEs) extracted from various resources tend to be heterogeneous. In fact, each source often has its own way to name the DE it uses. For instance, the DE for pathological conditions is named `Disorders` in GeneCards, but `Disease` in HPRD. We previously proposed to exploit knowledge from UMLS for resolving DE heterogeneities through linguistic approaches. We complete this work by exploiting a more general terminological resource, WordNet.

WORDNET

WordNet is organized into sets of synonymous terms (verbs, nouns, adjectives, and adverbs), called synsets, each of which representing one lexical concept. The database contains about 150,000 lexical items organized in over 115,000 synsets. Synsets are organized into a hierarchy. Ancestors and descendants are called respectively hypernyms and hyponyms in WordNet parlance. Version 2.1 is used in this study.

METHODS

Our method can be summarized as follows. Starting from the mapping of DEs to UMLS, we first perform a similar mapping to WordNet (WN). We then exploit WN properties to validate unique matches to UMLS and disambiguate multiple matches. Finally, we attempt to find indirect mappings to UMLS through WN.

Mapping DEs to WordNet. In order to map DEs to WN, we use the *wn* program[14] to associate terms with synsets. When the DE consists of more than one word, we map it to the longest spanning syntagm in WN. For instance, the DE `Mus Musculus` is mapped to the synset mus_musculus#n#1 rather to the two synsets mus#n#2 (type genus of the Muridae) and musculus#n#1 (muscle). When multiple matches are found in WN, we use the context of the synsets for disambiguation purposes. In practice, we favor synsets whose definition or hypernyms contain pre-defined keywords related to the biomedical domain (e.g. word bases such as biologic, medic, genetic, chromosom). For example, as shown in figure 1, the synset selected for the word "transcription" is the second one because of the presence of the biomedical term "genetics" in its definition.

[2] http://bioinformatics.weizmann.ac.il/cards/
[3] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
[4] http://genecards.weizmann.ac.il/geneloc/
[5] http://www.gene.ucl.ac.uk/nomenclature/
[6] http://www.hgmd.org/
[7] http://www.expasy.org/sprot/
[8] http://www.rcsb.org/pdb/
[9] http://www.hprd.org/
[10] http://www.ebi.ac.uk/interpro/
[11] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=omim
[12] http://www.informatics.jax.org/
[13] http://ghr.nlm.nih.gov/

[14] http://wordnet.princeton.edu/doc

- S: (n) transcription, written text (something written, especially copied from one medium to another, as a typewritten version of dictation)
- **S: (n) transcription ((genetics) the organic process whereby the DNA sequence in a gene is copied into mRNA; the process whereby a base sequence of messenger RNA is synthesized on a template of complementary DNA)**
- S: (n) transcription (a sound or television recording (e.g., from a broadcast to a tape recording))
- S: (n) arrangement, arranging, transcription (the act of arranging and adapting a piece of music)
- S: (n) recording, transcription (the act of making a record (especially an audio record))

Figure 1: Candidate synsets for the word "transcription"

Finally, we filter WN candidate synsets according to the grammatical form. For instance, in the DE `detailed genetic map`, the word "detailed" has three candidate synsets: one adjective and two verbs. Based on the syntactic analysis of the DE, only the adjective is selected here.
The mapping to WN results in a list of synsets, their definition, synonyms, and hypernyms associated with each DE.

Validating unique mappings to UMLS. Assuming a mapping to WN is found, this mapping itself is either unique or multiple. If the mapping to WN is unique, we exploit the properties of the candidate synset to validate the mapping to the UMLS. Toward this end, we compare the concept and synset according to the following criteria, in this order:
- Similarity of their definitions.
- Presence of common synonyms.
- Presence of common ancestors.

For criteria 2 and 3, we map the synonyms and hypernyms of the synset in WN to the UMLS through exact and normalized matches.
When several mappings to WN are found, this indicates that the synset is ambiguous or only partially represented in WN. In both cases, the mapping to WN is not useful for validating the mapping to UMLS. For example, the DE `Northern Blot` is mapped to the UMLS concept "Northern Blot" (Laboratory Procedure) and to the two WN synsets "northern" and "blot".

Disambiguating multiple mappings to UMLS. In order to disambiguate the multiple mappings of a DE to the UMLS, we map it to WN, resulting in one or more synsets for this DE. We then associate pairwise the UMLS concepts and WN synsets, respectively, and select the best (concept,synset) pair using the criteria described for the validation in the section above.

Identifying indirect mappings to UMLS through WordNet. For those DEs for which no mapping to UMLS concepts was found (i.e., when the only mapping candidates are WN synsets), we try to find an equivalent UMLS concept not from the DE itself, but from its mapping to WN. Starting from the synset(s) mapped to WN, we first attempt to map each of the synonyms in the synset(s) to the UMLS, using exact and normalized matches as before. If no synonym is mapped to UMLS, we start an equivalent mapping process for the direct hypernyms of the synset(s).
The resulting concepts constitute candidates for indirect mappings of the DE to UMLS through WN.

The whole process is illustrated in Figure 2.

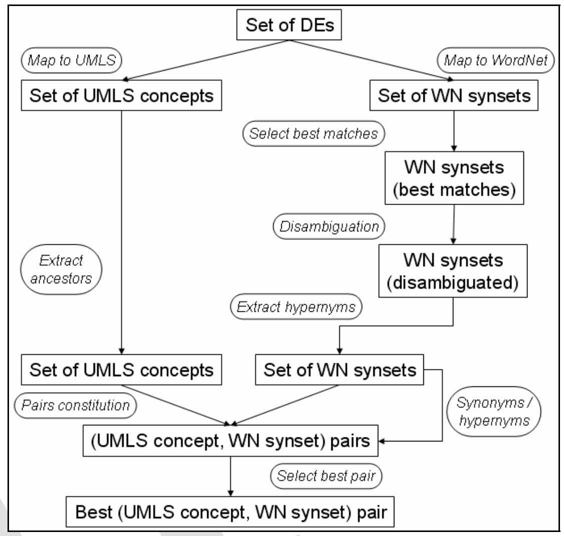

Figure 2: Direct and indirect mappings (through WordNet) of DEs to the UMLS

RESULTS

474 distinct DEs (548 tokens) were extracted from the eleven selected sources. Most of them were successfully mapped to WN. We provide the details of the mapping to WN with respect to the original mapping to UMLS and we evaluate the contribution of WN to improving the mapping of DEs to UMLS.

Mappings to UMLS vs. WordNet. The number of DEs in the three different categories of UMLS mappings are:
- Unique matches: 187 distinct DEs (39.5%).
- Multiple matches: 200 distinct DEs (42.1%).
- No matches: 87 distinct DEs (18.4%).

The results obtained for the two independent mappings to UMLS and WN, respectively, are presented in Table 1. The number and percentage of DEs mapped to these two resources are given, along with the number of distinct UMLS concepts and WN synsets mapped to DEs. The bottom row shows the results obtained by WN after disambiguation (Fig. 1) and elimination of ambiguous synsets when a unique concept exists (e.g., as with Northern Blot).

Table 1: Results obtained for the mapping of DEs to UMLS and to WordNet (independently)

	# DEs mapped	Unique matches	Multiple matches	No matches	# distinct concepts / synsets
UMLS	387 (82.1%)	187	200	87	458
WordNet	429 (90.5%)	105	324	45	1,878
After disambiguation	394 (83.1%)	259	135	80	558

Fifteen DEs (3.2%) were mapped to the UMLS only, including SNPs (Polymorphism, Single Nucleotide), rt-pcr (Reverse Transcriptase Polymerase Chain Reaction), and Micro-lesions. This finding is not surprising since these DEs are very specific to the biomedical domain.

Conversely, 55 DEs (11.6%) were mapped to WN only. Examples include Homology, Lineage, Products, Pathways, Transcripts, and Motifs.

Overall, 30 DEs (6.3%) were mapped to neither the UMLS, nor WN, including Paralogs, Ortholog, and Exuns.

Validating unique mappings to UMLS. WN provided supporting evidence for validating 82 unique mappings of DEs to UMLS (43.9%). More specifically, 68 were validated by exploiting definition similarity, 2 with synonyms, and 12 using ancestors. Following are some examples of mappings validated with respect to the type of evidence supporting the validation.

- The mapping of the concept "RNA, Messenger" (C0035696) to the DE mRNA sequence is validated by the synset mrna#n#1 because of the similarity in their definitions (51.9%). Common elements in definitions include "template for protein synthesis", "nucleus", and "RNA".
- The mapping of the concept "Duplication" (C0332597) to the DE Duplication is validated by the synset duplication#n#1 because they share a synonym: "Duplicate".
- The mapping of the concept "Length" (C1444754) to the DE Length is validated by the synset length#n#1 because they share common ancestors, "Dimensions" (C0439534) and "Attribute" (C0449234).

73 cases (39.0%) could not be validated by mapping to WN because their mapping to WN was ambiguous. For example, the DE Gene Function was mapped to only one UMLS concept "Gene Function" (C0314627), but to four synsets in WN.

Finally, 32 unique mappings (17.1%) could not be validated because WN properties did not permit to find common characteristics between the concept and synset associated with the DE.

Disambiguating multiple mappings to UMLS. 95 multiple mappings of DEs to UMLS (47.5%) were successfully disambiguated with WN. Nearly all of them resulted from processing the definitions (94 compared to only one for the ancestors). An example is the mapping of the DE Protein. Initially, it resulted in three concepts: "Protein" (C0033684), "Protein measurement" (C0202202), and "Protein location" (C1325816). Through the mapping to the synset protein#n#1, we were able to select the concept "Protein" because of the similarity in their definitions (34.8%).

31 multiple mappings of DEs to UMLS (15.5%) were not disambiguated because there was no associated match in WN. The remaining 75 mappings (37.0%) could not be disambiguated because there was more than one WN candidate synset or no proposed (concept,synset) pair could be selected as the best one.

Identifying indirect mappings to UMLS through WordNet. Overall, 37 additional indirect mappings of DEs to UMLS (42.5%) were identified through WN. 10 were valid, 26 ambiguous, 1 erroneous.

By exploiting synonymy in WN, 16 indirect mappings of DEs to UMLS were suggested. For instance, no direct mapping to the UMLS was identified for the DE topology, because no UMLS concept has topology as a synonym. However, this DE is mapped to the synset topology#n#2, of which one synonym is "regional anatomy". Unlike topology, "regional anatomy" can be mapped to the UMLS (concept C0002812). The DE topology can thus be mapped to the UMLS concept "Regional anatomy" (C0002812), through a synonym from WN.

Using direct ancestors in WN, 21 indirect mappings to UMLS were found. An example is the DE Product which is mapped to the synset product#n#4. No synonym exists in this synset, but its direct hypernym "Chemical" is a UMLS concept (C0220806), which thus constitutes a potential UMLS mapping of the DE Product.

Example. In order to illustrate the contribution of WN, we describe the mapping to the UMLS of the DE *Transcription data* extracted from the source GeneCards (Figure 3). In the UMLS, a partial match is found to "Transcription" (concept C0040649). In WN, two partial matches are found: "Transcription" to five synsets and "data" to the synsets data#n#1 and data#n#2. The disambiguation process of "Transcription" is shown in Figure 1, leading to select the synset transcription#n#2. The synset data#n#1 is chosen over data#n#2 because of the presence of its synonym "information" in the set of DEs (context). From the two independent mappings, we now can:

- i) confirm that the mapping to the concept "Transcription, Genetic " (C0040649) is correct according to similarity in the definitions of C0040649 and transcription#n#2.

- ii) propose an indirect mapping of the word “data” to the concept “Information” (C0205549), through the synset data#n#1 which maps to the original DE and has a synonym in UMLS (C0040649).

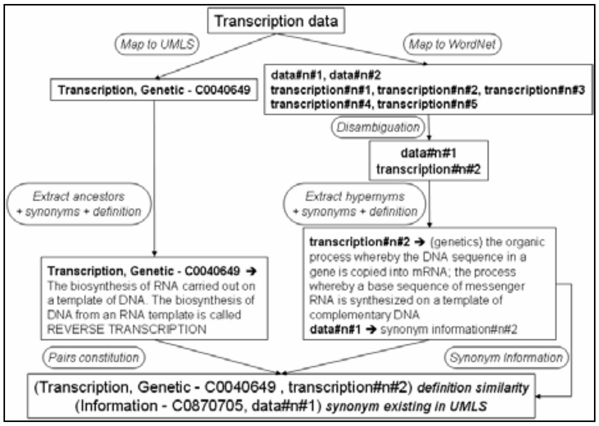

Figure 3: Example of the mapping process for the DE `Transcription data`

DISCUSSION

WordNet contribution

Overall, for the 474 studied DEs, WN contributed to validate 82 mappings and disambiguate 95 mappings and to identify 37 indirect mappings of DEs to the UMLS, illustrating its effective contribution.

As shown in the results, the exploitation of synonyms in our method was generally not useful. In fact, only two mappings could be validated using synonyms. This can probably be explained by the small number of synonyms present in WN, especially compared to large terminological systems such as the UMLS.

Another finding is the relatively low similarity between some definitions. It is simply due to the fact that some UMLS definitions are very long (cf. Fig. 3), resulting in a small percentage of common words. Indeed, the 34.8% of similarity between “Protein” and protein#n#1 may seem low, but the concept and the synset still share five relevant elements (“organic”, “group”, “amino acids”, “living cells”, and “polymer”). This information, however, is sufficient to select the UMLS concept “Protein” over the two other candidate concepts.

Most of indirect mappings proposed by WN are ambiguous (70.3%). For instance, the DE `contributor`, which is not mapped directly to the UMLS, is mapped to two synsets: contributor#n#1, which has “Writer” and “Author” whose direct hypernyms exist in UMLS (C0341628 and C0221192). Conversely, contributor#n#2 has the direct ancestor “Donor” (C0013018), which is also found in the UMLS. In this case, a manual review is necessary to select which one, if any, of the proposed indirect mapping is correct.

Future work

Indirect mappings of DEs to UMLS through their values. Some DEs remain unmapped to the UMLS even through synonyms and hypernyms. We plan to define an alternative approach to mapping DEs to the UMLS, which consists in mapping not the DEs themselves to WN, but their associated values. For example, the DE `homology` present in Entrez Gene is found in WN (synset homology#n#1) but can not be mapped to the UMLS. However, its values include “Mouse”, “Rat”, and “Human” indicating that this DE gives information about organisms (among which some variant of a gene is shared). Analysing these DE values could allow to associate the DE `homology` to the DE `Organism` existing in Swiss-Prot. It is important to notice that the use of WN is essential in these cases since it enables to represent a DE (with one of its synset), which would have been ignored if only UMLS mapping had been achieved.

Exploiting structural properties. We use the structural properties of the UMLS and WN to validate and disambiguate the mappings of DEs to the UMLS as well as to identify new mappings. Mork also used structural properties to align representations of anatomy [7]. In our study, the exploitation of ancestors was useful to validate 12 original mappings to the UMLS. Moreover, it provided 21 new (indirect) mappings of DEs to the UMLS. Our method is however limited to ancestors. Analogously, we want to exploit descendants to search for additional entities that are common to UMLS and WN.

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

1. Mougin F, Burgun A, Bodenreider O. Data integration through data elements: Mapping data elements to terminological resources. Proc Symp on Semantic Mining in Biomedicine. 2006
2. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med; 1993;32(4):281-291
3. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care. 1994;:235-9
4. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21
5. Miller GA. WordNet: A Lexical Database for English. ACM Communications , 38(11), Nov. 1995
6. Burgun A, Bodenreider O. Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. Proc Workshop on WordNet and Other Lexical Resources. 2001
7. Mork P, Bernstein PA. Adapting a Generic Match Algorithm to Align Ontologies of Human Anatomy. International Conference on Data Engineering. 2004: 787-790